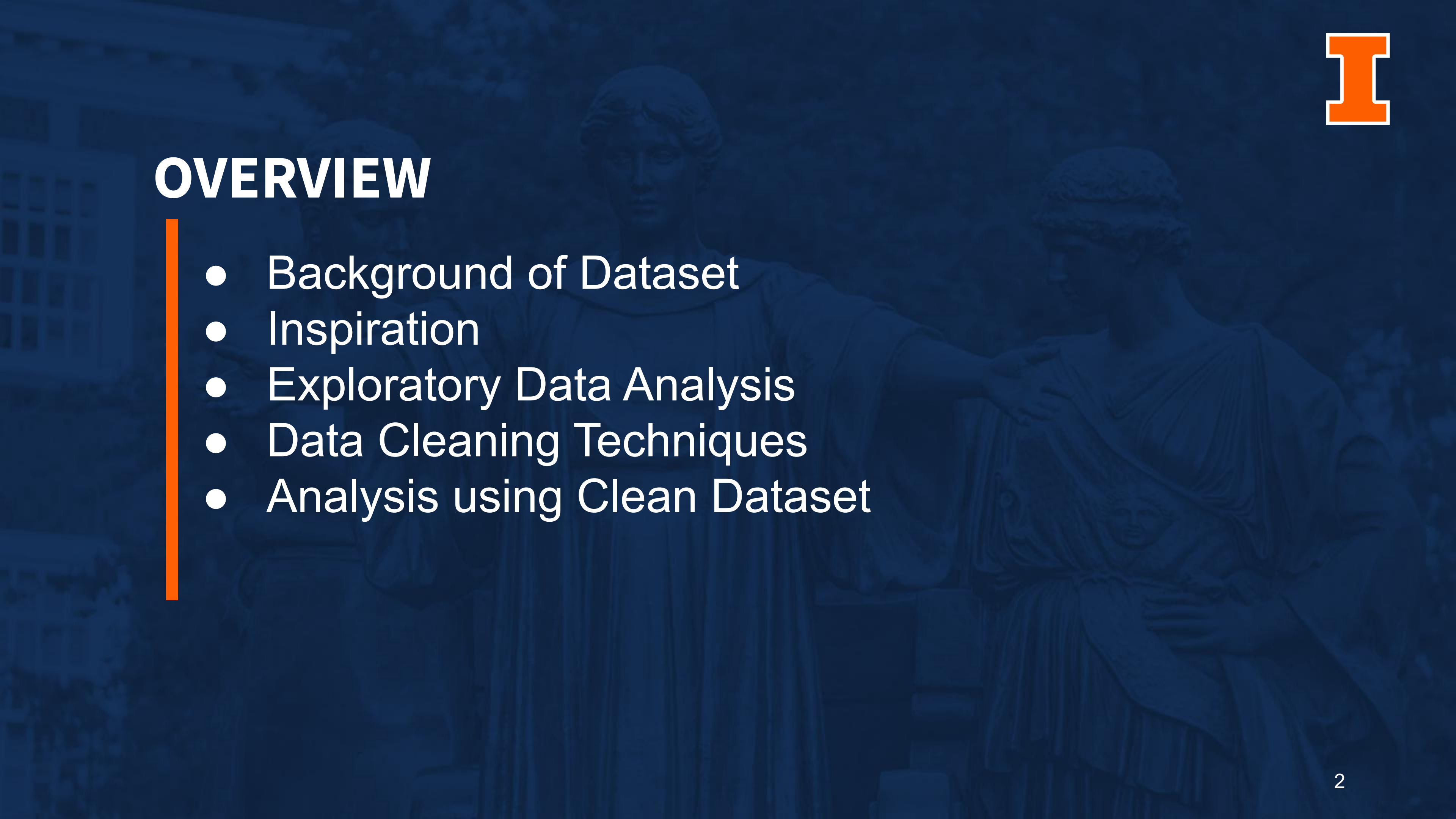
Final Project Presentation

IS537: Theory and Practice of Data Cleaning

Presented By:

Adit Rathi
Ashwini Karkhanis
Himani Mehta



OVERVIEW

- Background of Dataset
- Inspiration
- Exploratory Data Analysis
- Data Cleaning Techniques
- Analysis using Clean Dataset

Background of Dataset

Link: https://www.kaggle.com/datasets/dgomonov/new-york-city-airbnb-open-data

- Airbnb provides various rental options for different customer segments
- $75 Billion online marketplace for renting out homes/villas/ private rooms
- The data can be analyzed and used for security, business decisions, understanding of customers' and providers' (hosts) behavior,etc
- Dataset has information about hosts, geographical availability, necessary metrics to make predictions and draw conclusions

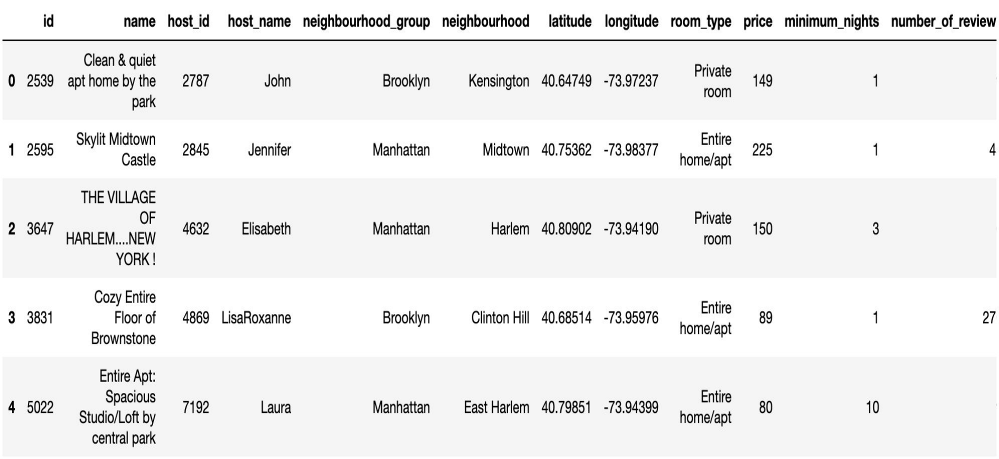
Inspiration

Questions that can be answered by the dataset:

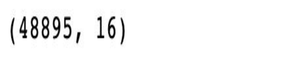
- Top neighbourhoods in NYC with respect to average price/day of Airbnb listings?
- How do monthly reviews vary with room types in each neighbourhood groups?
- Room_types vs price on different neighbourhood groups?
- Which neighborhood has the highest number of properties?
- On average how many nights people stayed in each room_types?

Exploratory Data Analysis

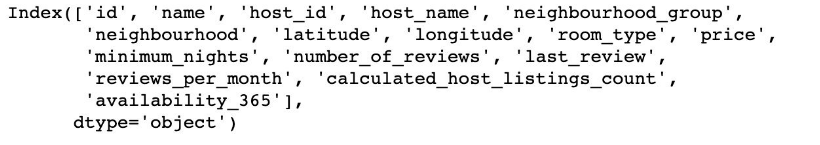
1. airbnb.head()
2. airbnb.shape()
3. airbnb.columns()



	id	name	host_id	host_name	neighbourhood_group	neighbourhood	latitude	longitude	room_type	price	minimum_nights	number_of_review
0	2539	Clean & quiet apt home by the park	2787	John	Brooklyn	Kensington	40.64749	-73.97237	Private room	149	1	
1	2595	Skylit Midtown Castle	2845	Jennifer	Manhattan	Midtown	40.75362	-73.98377	Entire home/apt	225	1	4
2	3647	THE VILLAGE OF HARLEM....NEW YORK !	4632	Elisabeth	Manhattan	Harlem	40.80902	-73.94190	Private room	150	3	
3	3831	Cozy Entire Floor of Brownstone	4869	LisaRoxanne	Brooklyn	Clinton Hill	40.68514	-73.95976	Entire home/apt	89	1	27
4	5022	Entire Apt: Spacious Studio/Loft by central park	7192	Laura	Manhattan	East Harlem	40.79851	-73.94399	Entire home/apt	80	10	



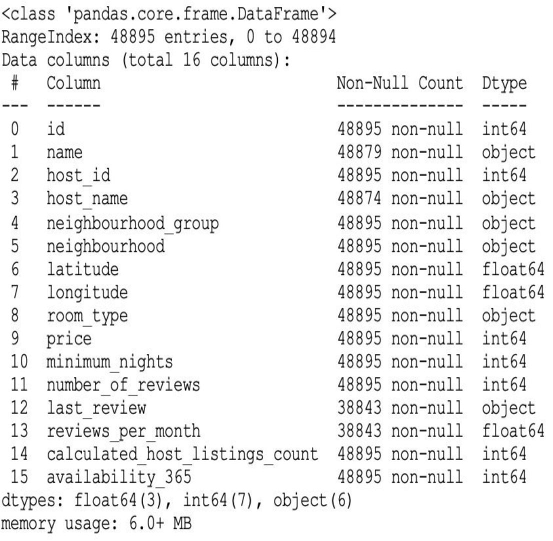
```
(48895, 16)
```



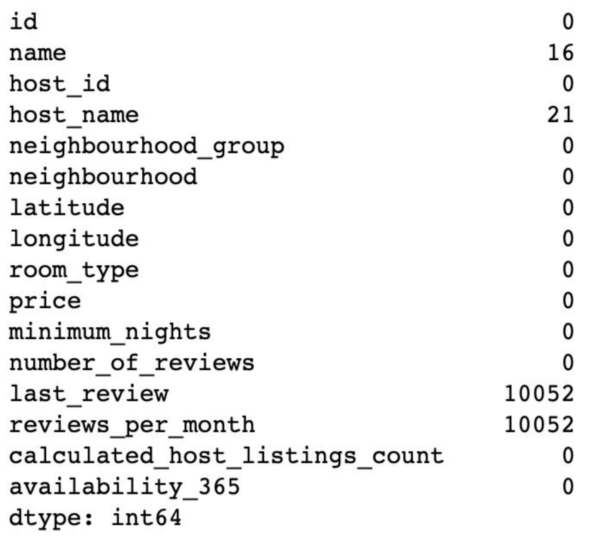
```
Index(['id', 'name', 'host_id', 'host_name', 'neighbourhood_group',
       'neighbourhood', 'latitude', 'longitude', 'room_type', 'price',
       'minimum_nights', 'number_of_reviews', 'last_review',
       'reviews_per_month', 'calculated_host_listings_count',
       'availability_365'],
      dtype='object')
```

Exploratory Data Analysis

4. airbnb.info()
5. airbnb.describe()
6. airbnb.isnull().sum()



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 48895 entries, 0 to 48894
Data columns (total 16 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   id                              48895 non-null  int64
 1   name                            48879 non-null  object
 2   host_id                         48895 non-null  int64
 3   host_name                       48874 non-null  object
 4   neighbourhood_group             48895 non-null  object
 5   neighbourhood                   48895 non-null  object
 6   latitude                        48895 non-null  float64
 7   longitude                       48895 non-null  float64
 8   room_type                       48895 non-null  object
 9   price                           48895 non-null  int64
 10  minimum_nights                  48895 non-null  int64
 11  number_of_reviews               48895 non-null  int64
 12  last_review                     38843 non-null  object
 13  reviews_per_month               38843 non-null  float64
 14  calculated_host_listings_count  48895 non-null  int64
 15  availability_365                48895 non-null  int64
dtypes: float64(3), int64(7), object(6)
memory usage: 6.0+ MB
```



```
id                                    0
name                                 16
host_id                               0
host_name                            21
neighbourhood_group                   0
neighbourhood                         0
latitude                              0
longitude                             0
room_type                             0
price                                 0
minimum_nights                        0
number_of_reviews                     0
last_review                       10052
reviews_per_month                 10052
calculated_host_listings_count        0
availability_365                      0
dtype: int64
```

Data Cleaning Techniques

Dealing with NULL values

1. Name and host_name
   Columns dropped as it cannot be retrieved

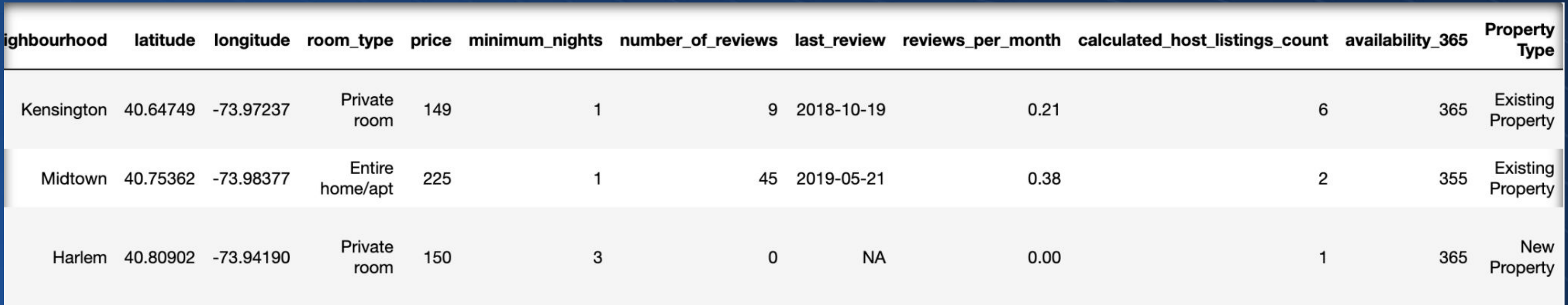
2. last_review
   Replaced all NaN values with NA

3. reviews_per_month
   Replaced all NaN with 0

Data Cleaning Techniques

4. Adding new column Property Type
Properties where the number_of_reviews is 0, last_review and reviews_per_month are only NULL values

Dataset after adding new column:



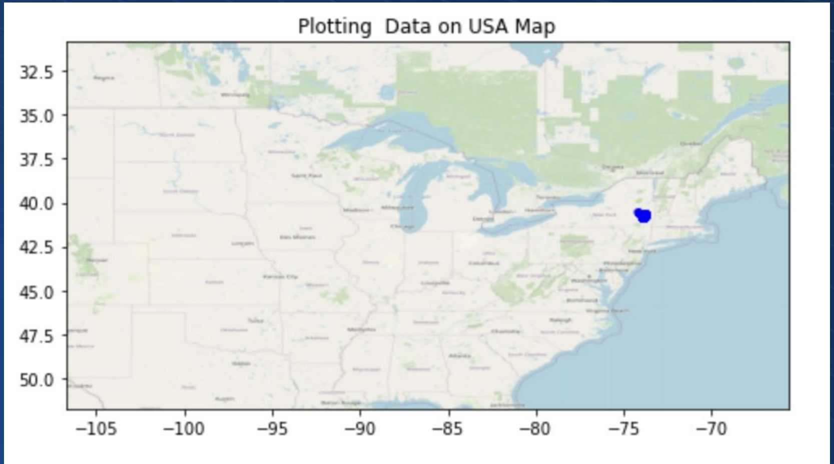
ighbourhood	latitude	longitude	room_type	price	minimum_nights	number_of_reviews	last_review	reviews_per_month	calculated_host_listings_count	availability_365	Property Type
Kensington	40.64749	-73.97237	Private room	149	1	9	2018-10-19	0.21	6	365	Existing Property
Midtown	40.75362	-73.98377	Entire home/apt	225	1	45	2019-05-21	0.38	2	355	Existing Property
Harlem	40.80902	-73.94190	Private room	150	3	0	NA	0.00	1	365	New Property

Data Cleaning Techniques

<u>Checking and Dealing with Outliers</u>

1. Location

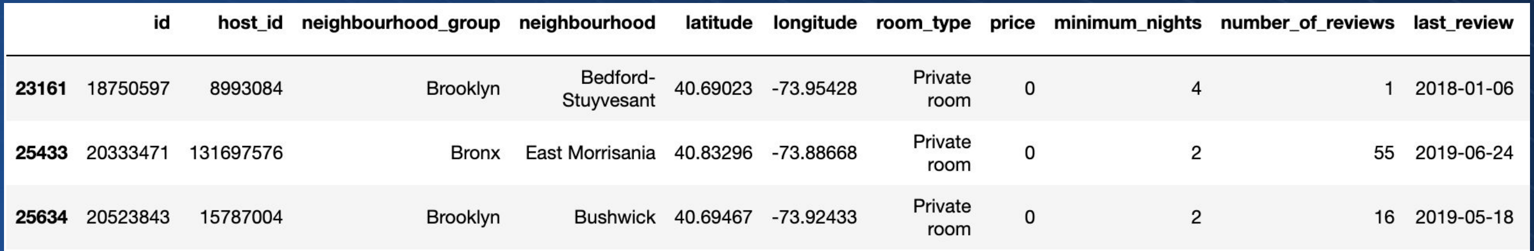
   The longitude and latitude were plotted to check if all the locations belong to New York city



Data Cleaning Techniques

6. Replacing with average value
Properties with price 0 replaced with average price of all the properties belonging to the same neighborhood

Properties with price as 0

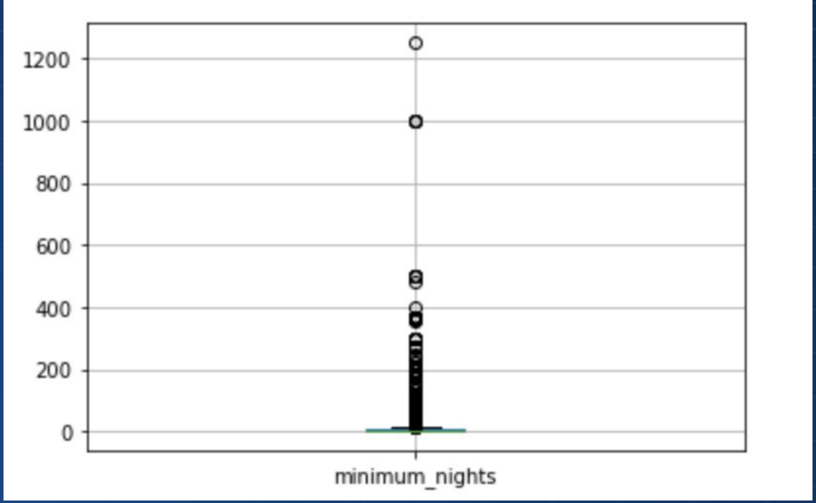


	id	host_id	neighbourhood_group	neighbourhood	latitude	longitude	room_type	price	minimum_nights	number_of_reviews	last_review
23161	18750597	8993084	Brooklyn	Bedford-Stuyvesant	40.69023	-73.95428	Private room	0	4	1	2018-01-06
25433	20333471	131697576	Bronx	East Morrisania	40.83296	-73.88668	Private room	0	2	55	2019-06-24
25634	20523843	15787004	Brooklyn	Bushwick	40.69467	-73.92433	Private room	0	2	16	2019-05-18

Data Cleaning Techniques

2. minimum_nights

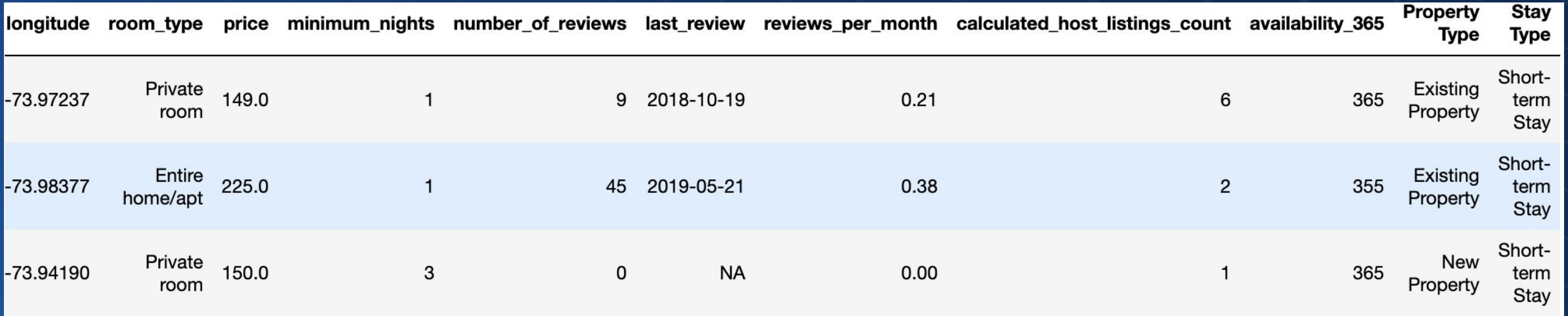
As per Airbnb policy, the stay duration in a single booking cannot exceed 90 days. So, replaced all listing having minimum_nights above 90 with 90.



Data Cleaning Techniques

5. Adding new column Stay Type
Marking all properties with minimun_nights above 28 days as Long-term Stay and below 28 days as Short-term stay

Dataset after adding new column:



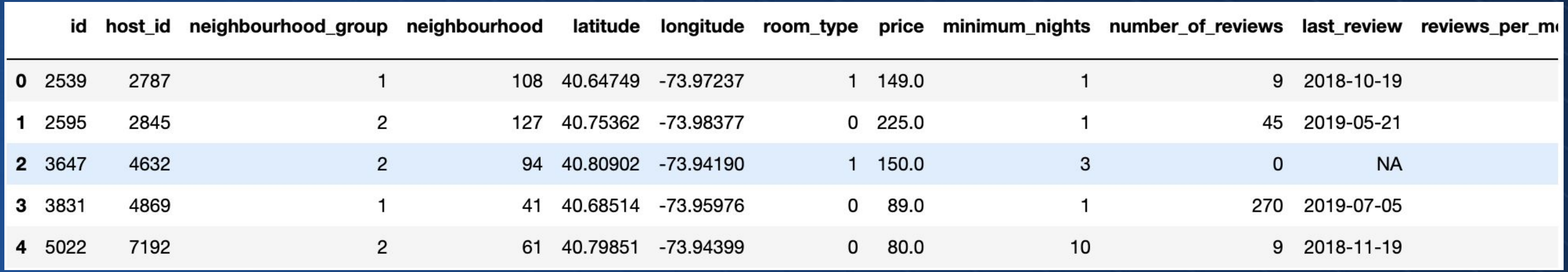
longitude	room_type	price	minimum_nights	number_of_reviews	last_review	reviews_per_month	calculated_host_listings_count	availability_365	Property Type	Stay Type
-73.97237	Private room	149.0	1	9	2018-10-19	0.21	6	365	Existing Property	Short-term Stay
-73.98377	Entire home/apt	225.0	1	45	2019-05-21	0.38	2	355	Existing Property	Short-term Stay
-73.94190	Private room	150.0	3	0	NA	0.00	1	365	New Property	Short-term Stay

Data Cleaning Techniques

Label Encoding of categorical variables

Performed on columns having categorical data and convert it into machine readable format

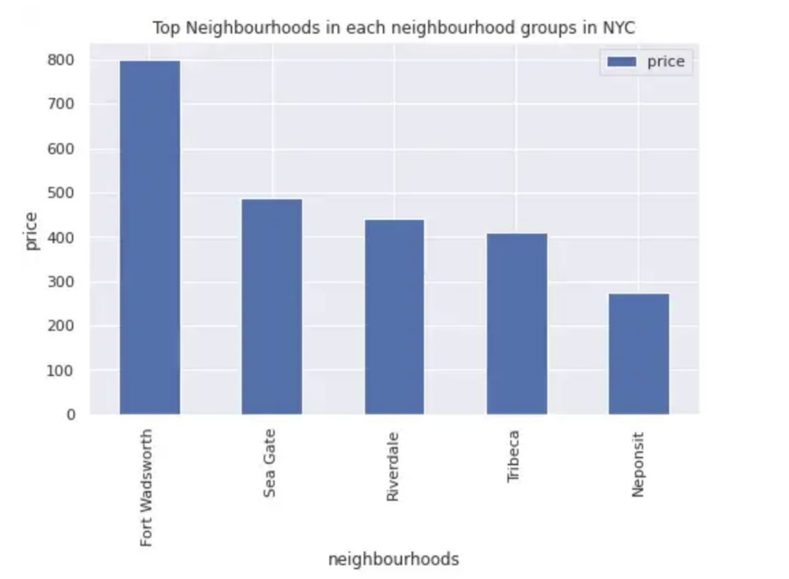
Columns: room_type, neighbourhood_group, neighbourhood, Property Type, Stay Type



	id	host_id	neighbourhood_group	neighbourhood	latitude	longitude	room_type	price	minimum_nights	number_of_reviews	last_review	reviews_per_m
0	2539	2787	1	108	40.64749	-73.97237	1	149.0	1	9	2018-10-19	
1	2595	2845	2	127	40.75362	-73.98377	0	225.0	1	45	2019-05-21	
2	3647	4632	2	94	40.80902	-73.94190	1	150.0	3	0	NA	
3	3831	4869	1	41	40.68514	-73.95976	0	89.0	1	270	2019-07-05	
4	5022	7192	2	61	40.79851	-73.94399	0	80.0	10	9	2018-11-19	

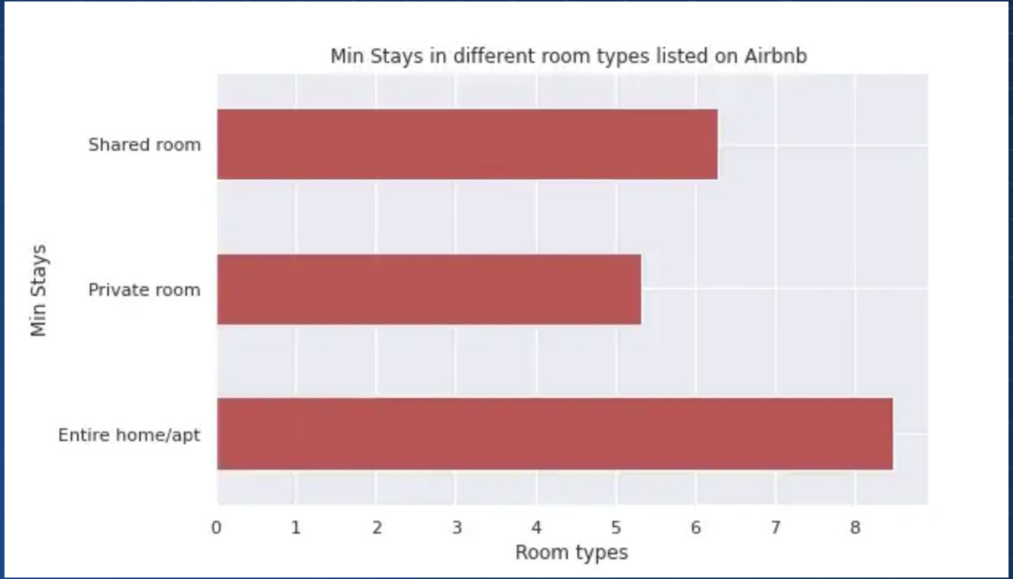
Analysis using Clean Data

1. Top neighbourhoods in NYC with respect to average price/day of Airbnb listings?

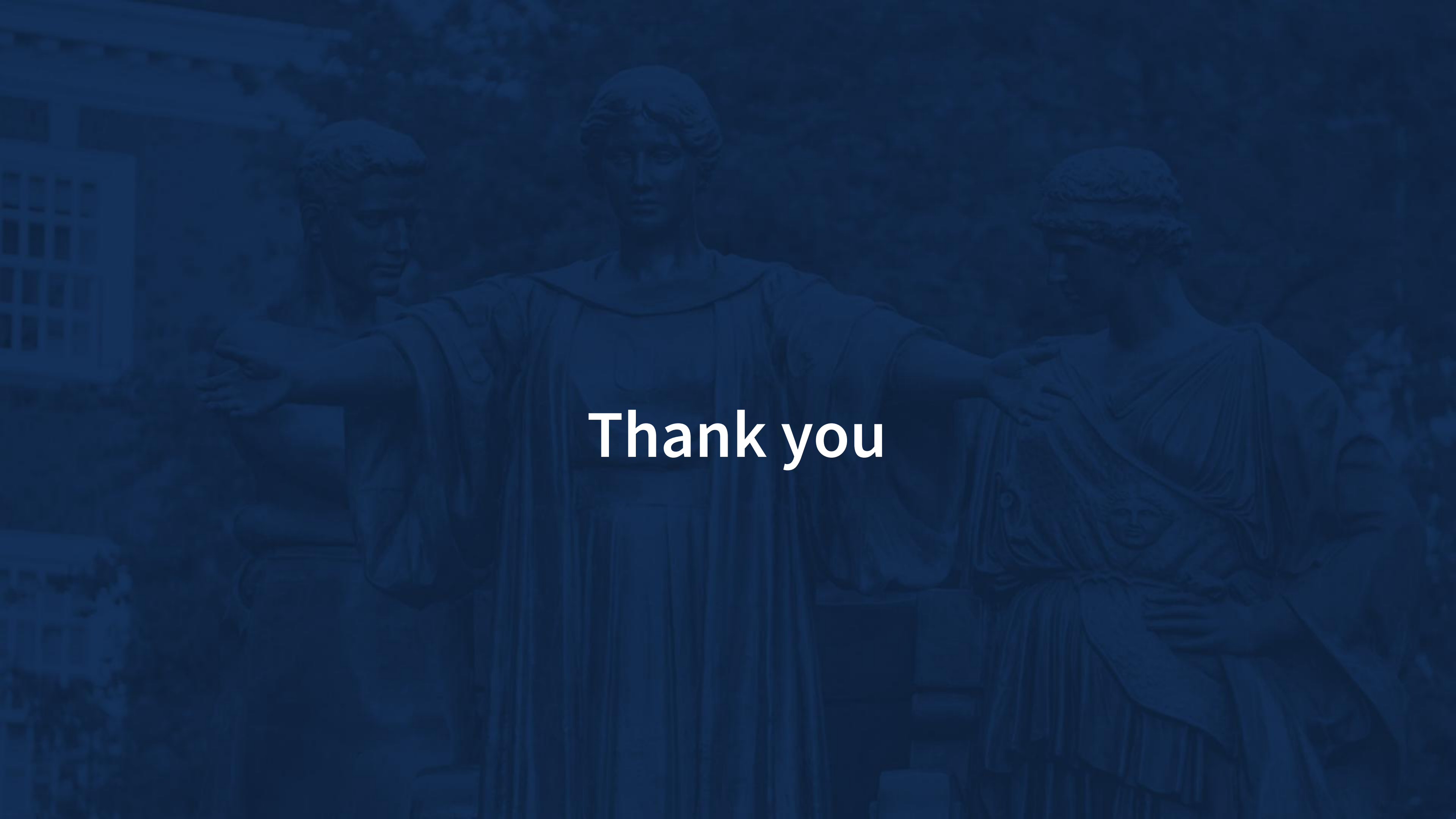


Analysis using Clean Data

1. On an average for how many nights people stayed in each room_types?



Questions?



Thank you